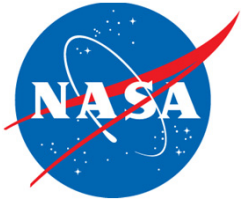
Machine Learning Methods for Real Time and Archival Classification of Astronomical Transients and Variables



Umaa Rebbapragada
Machine Learning and Instrument Autonomy
Jet Propulsion Laboratory, California Institute of Technology

March 1, 2012
Center for Time Domain Informatics
University of California, Berkeley, CA

Agenda

- Overview of Research Interests
- ASKAP / VAST
  - Source type classification
  - Results of study on simulated data
  - Archival, real time classification
- Palomar Transient Factory
  - Binary real time classification
  - Preliminary results

Agenda

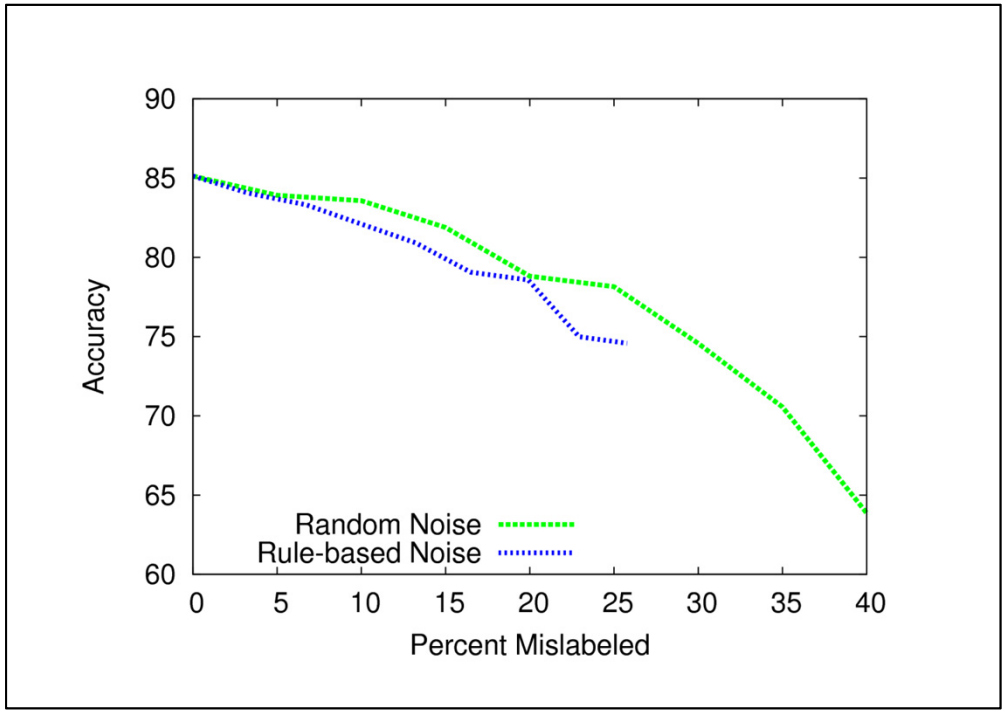
- **Overview of Research Interests**

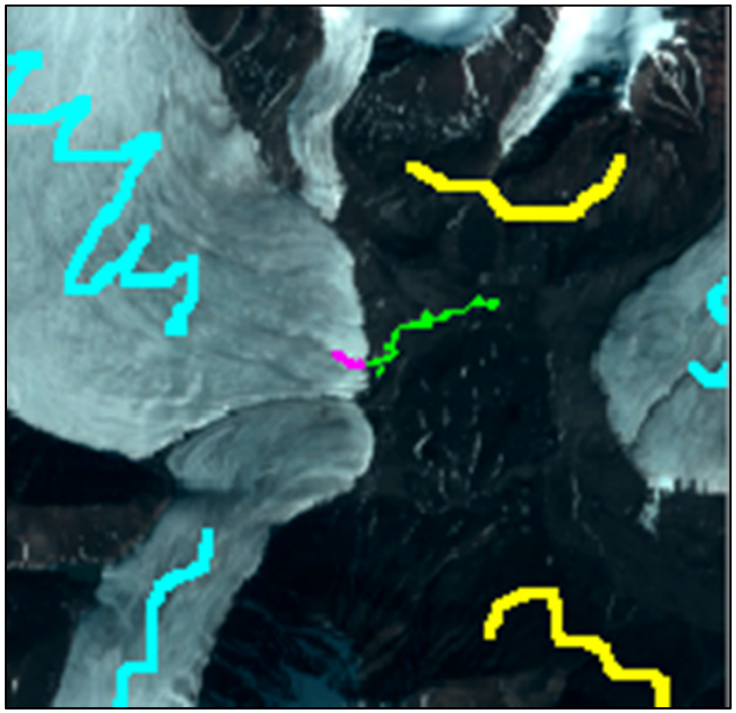
- ASKAP / VAST
  - Source type classification
  - Results of study on simulated data
  - Archival, online classification

- Palomar Transient Factory
  - Binary real time classification
  - Preliminary results

Mislabeled Training Data

- Mitigating effects of mislabeled training data
- Applied to landcover classification





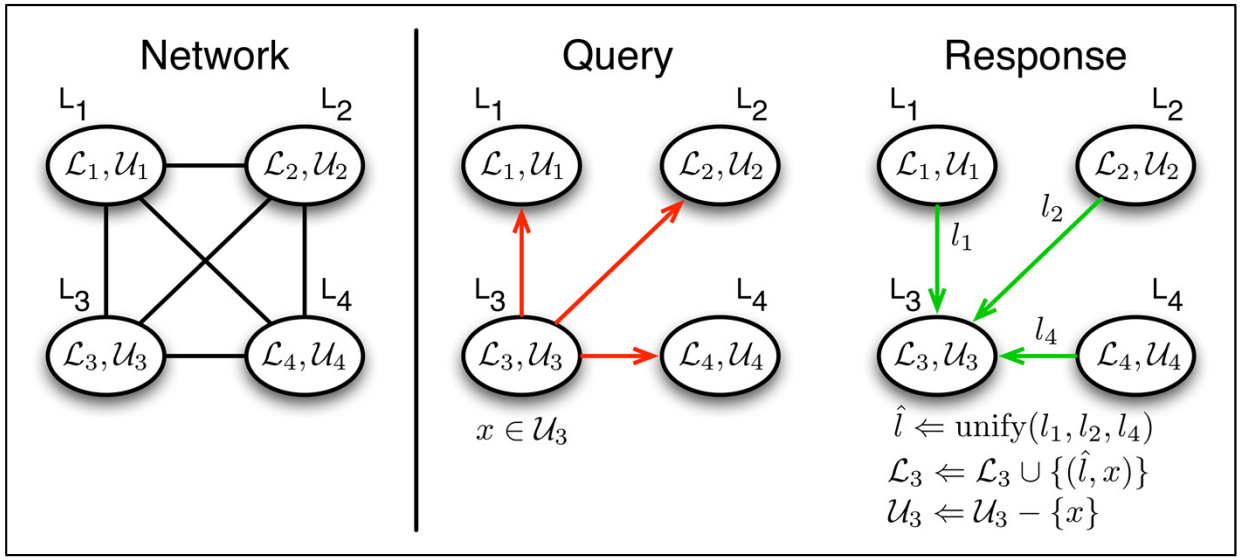
Key Publications

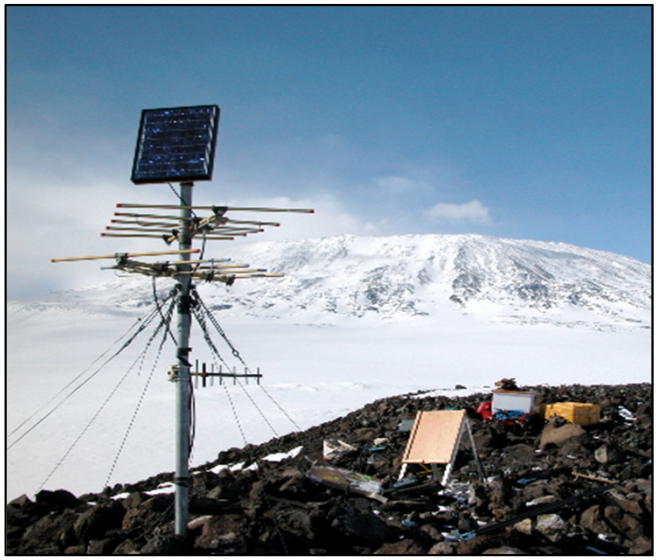
- **Detecting Mislabeled Training Data**

  - **U. Rebbapragada**, C. E. Brodley, D. Sulla-Menashe, M. Friedl (2012) Active Label Correction. Submitted to KDD 2012.
  - **U. Rebbapragada**, L. Mandrake, K. Wagstaff, D. Gleeson, R. Castano, S. Chien, and C. E. Brodley (2009) Improving Onboard Analysis of Hyperion Images by Filtering Mislabeled Training Data Examples. In *Proceedings of the 2009 IEEE Aerospace Conference*
  - **U. Rebbapragada**, R. Lomasky, C. E. Brodley and M. Friedl (2008) Generating High-Quality Training Data for Automated Land-Cover Mapping. In *Proceedings of the 2008 IEEE International Geoscience and Remote Science Symposium.*
  - **U. Rebbapragada** and C. E. Brodley (2007) Class Noise Mitigation Through Instance Weighting. In *Proceedings of the 18th European Conference on Machine Learning*

Multi-view Learning

- Semi-supervised learning from networked sensor data
- Motivated by ground sensors at Mount Erebus Volcanic Observatory





Other Research Interests

- Learning from crowdsourcing to identify earthquake-induced damages from satellite and aerial photography
- Tracking moving targets in aerial photography
- Anomaly Detection of Periodic Time Series
  - OGLE data

Key Publications

- **Collaborative Learning of Sensor Networks**
  - **U. Rebbapragada** and Kiri L. Wagstaff. Using Ensemble Decisions and Active Selection to Improve Low-Cost Labeling for Multi-View Data**.** *Proceedings of the ICML Workshop on Combining Learning Strategies to Reduce Label Cost*, July 2011.
- **Crowdsourced Learning of Earthquake-Induced Damage Assessment**
  - **U. Rebbapragada** and Thomas Oommen. Integrating Machine Learning into a Crowdsourced Model for Earthquake-Induced Damage Assessment. *Proceedings of the ICML Workshop on Machine Learning for Global Challenges*, July 2011
- **Anomaly Detection in Periodic Time Series**
  - **U. Rebbapragada**, P. Protopapas, C. E. Brodley and C. Alcock (2009) Finding Anomalous Periodic Time Series: An Application to Catalogs of Periodic Variable Stars. *Machine Learning*, Vol. 74, Issue 3, p. 281

Agenda

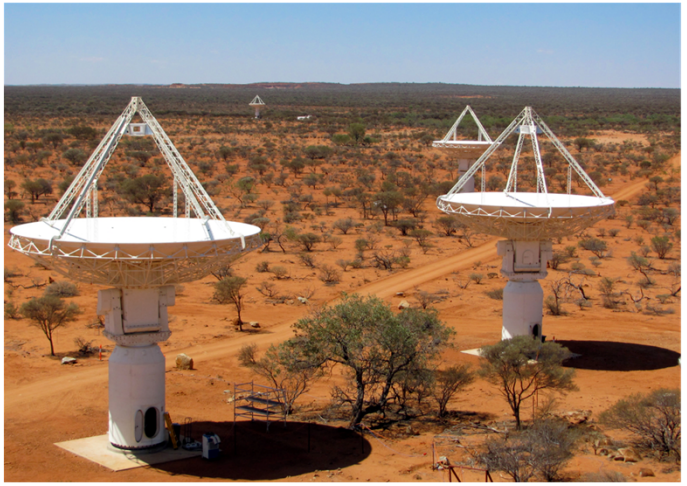
- Overview of Research Interests
- **ASKAP / VAST**
  - Source type classification
  - Results of study on simulated data
  - Archival, online classification
- Palomar Transient Factory
  - Binary real time classification
  - Preliminary results

Agenda

- Overview of Research Interests
- **ASKAP / VAST**
  - Source type classification
  - Results of study on simulated data
  - Archival, online classification
- Palomar Transient Factory
  - Binary real time classification
  - Preliminary results

ASKAP and VAST

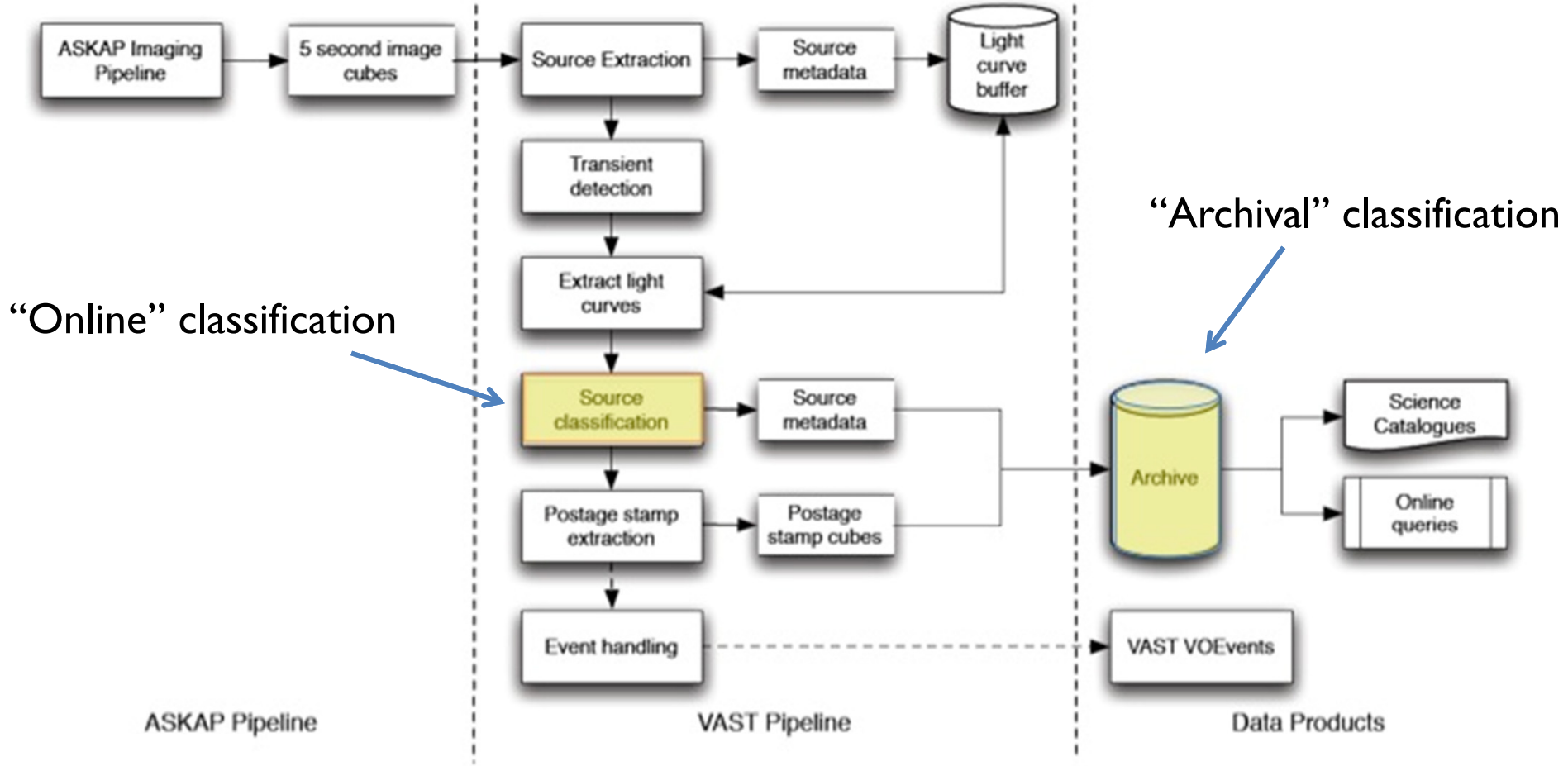
- **A**ustralian **SKA P**athfinder
- Observes radio sky in single day
- sub-mJy sensitivity
- 5 second cadence
- ASKAP BETA online in 2012



CSIRO's ASKAP antennas at the MRO in Western Australia. Credit: Antony Schinckel, CSIRO.

- **V**ariables **a**nd **S**low **T**ransients
- Potential to discover new objects and object classes
- Detection in real time

VAST Data Processing Pipeline

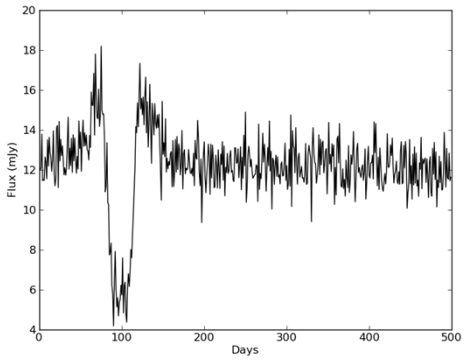


Study Goals

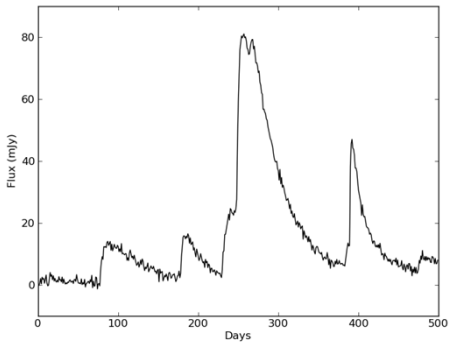
- Simulate VAST light curves
- Identify
  - feature representations
  - learning algorithms
- Estimate archival classification performance
  - by observing strategy
  - for source types of interest
- Estimate online classification performance
- **U. Rebbapragada**, K. Lo, K. Wagstaff, C. Reed and T. Murphy (2012) "VAST Memo #5: Offline and Online Classification of Simulated VAST Transients"
- http://www.physics.usyd.edu.au/sifa/vast/index.php/Main/Documents

Simulated Radio Source Types

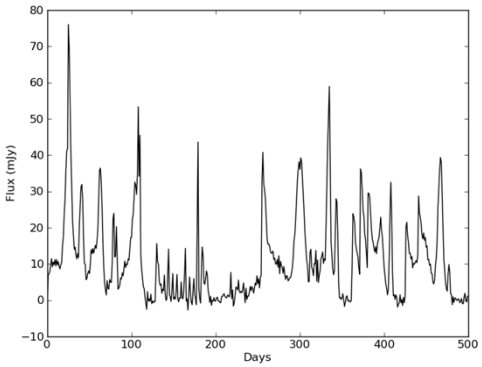
- 200 per source type, sampled daily for 400 days
- SNR at 3, 5, 7, 10σ



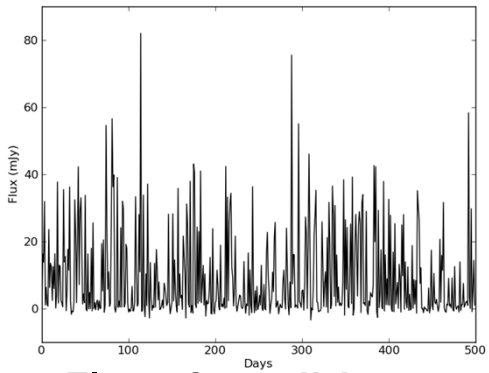
Extreme Scattering Event (ESE)



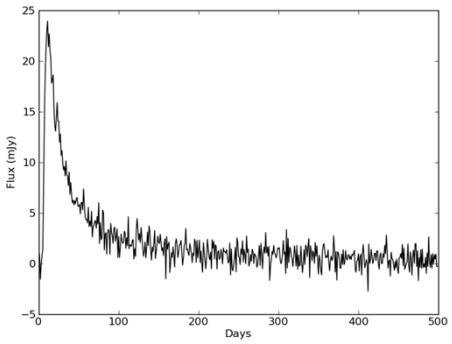
Xray Binary (XRB)



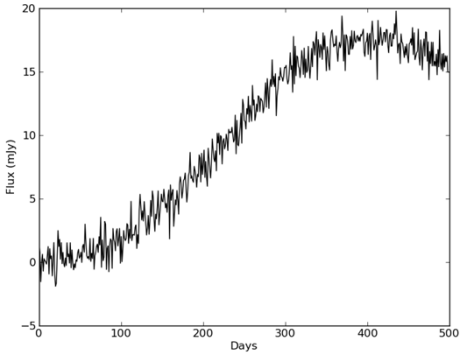
Flare Stare RSCVn



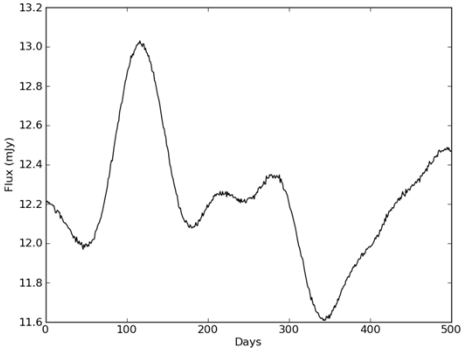
Flare Star dMe



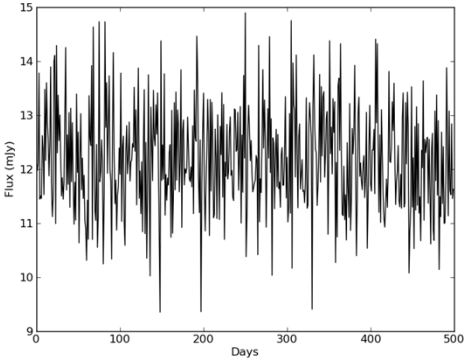
Supernova (SNe)



Nova



Intraday Variable (IDV)



Background Source (BG)

Observational Strategies

	VAST Wide	VAST Galactic Plain	VAST Deep	patches	monthly	log
Sampling	Daily	Irregularly sampled, at least once per week	Days 1, 2, 3, 4, 17, 21	Random 3 conse-cutive days per month	Every 30 days	Days 1, 2, 4, 8, etc.
RMS	0.5mJy	**0.1mJy**	**0.05mJy**	0.5mJy	0.5mJy	0.5mJy

Learning Algorithms

- **Support Vector Machine**
  - RBF kernel
- **Decision Tree**
- **Random Forest**
  - Ensemble of 10 unpruned decision trees
- Naïve Bayes
- Logistic Regression

Light Curve Characterization

- Frequency Domain
  - Lomb-Scargle Periodogram (**lsp**)
  - Haar Wavelets (**wlet**)

- Statistical Representations (**stat**)
  - Non-periodic features from [1]
  - Moment, shape statistics

[1] Richards et al. (2011) On machine- learned classification of variable stars with sparse and noisy time-series data. arXiv 1101.1959

Feature Combinations

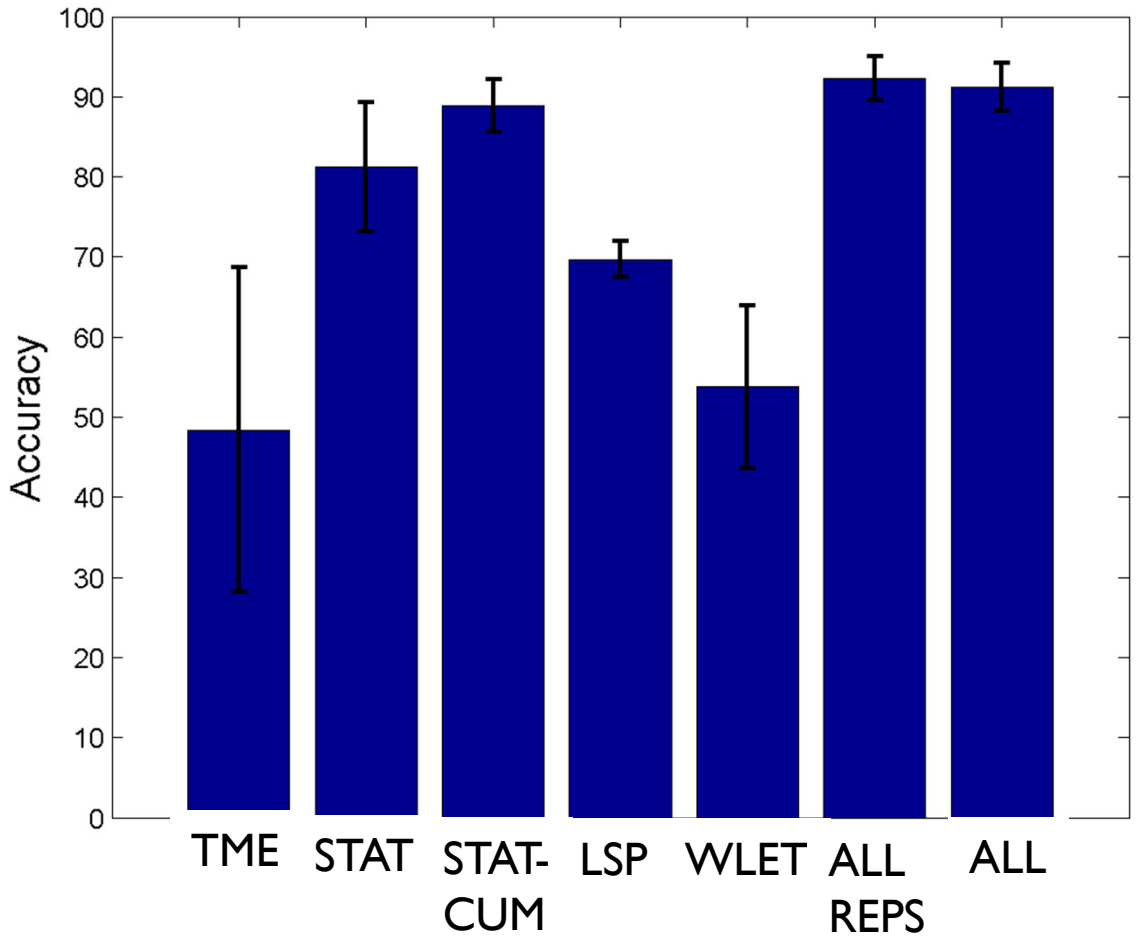
- Cumulative Statistics (stat-cum):
  - Extract statistical features after k, 2k, 3k observations

- Concatenate:
  - All representations: lsp + wlet + stat-cum
  - Everthing: time + lsp + wlet + stat-cum

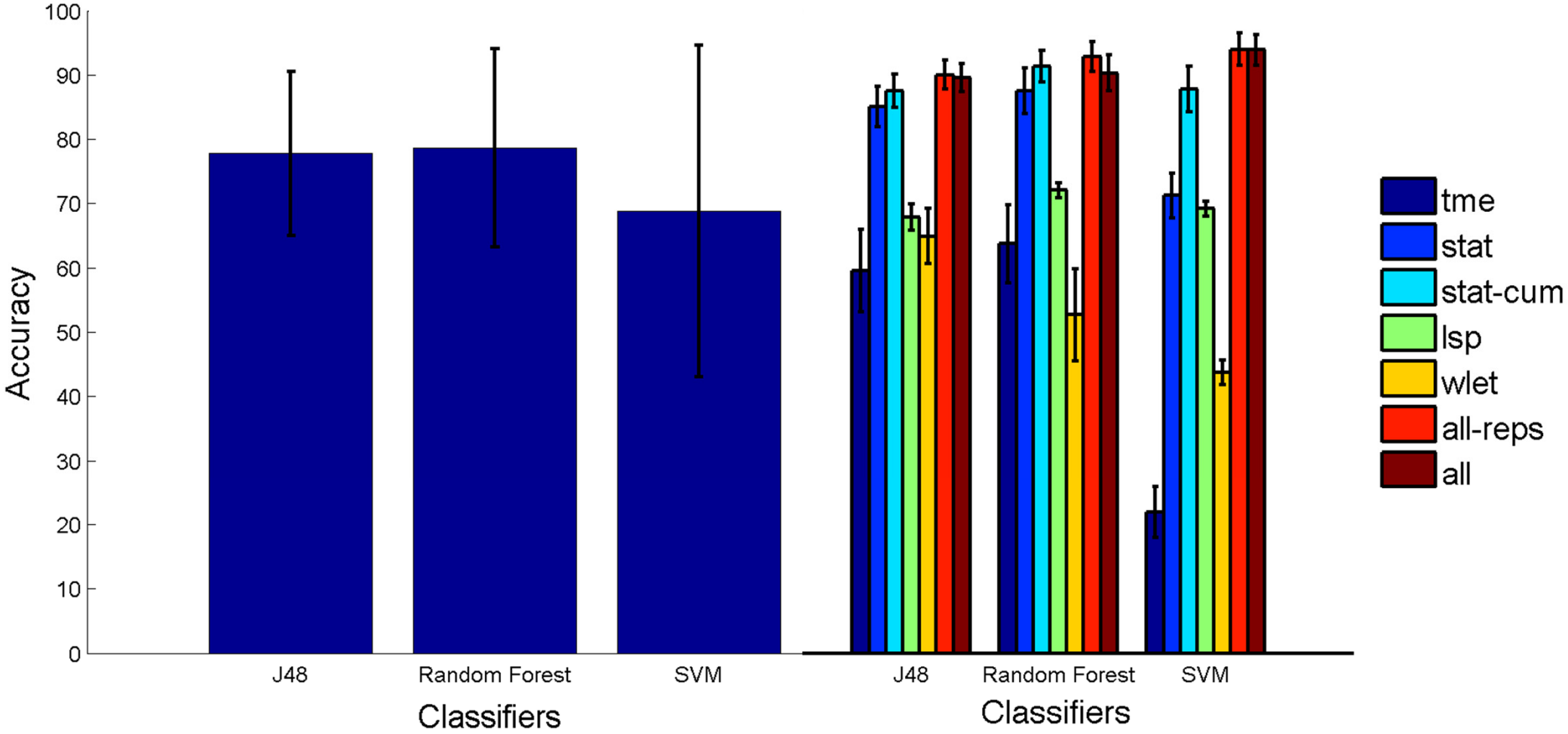
Accuracy by Feature

- VAST Wide
- Averaged over all SNR, classifiers
- 10-fold CV

- **Stat-cum better than stat**
- **All-reps best**

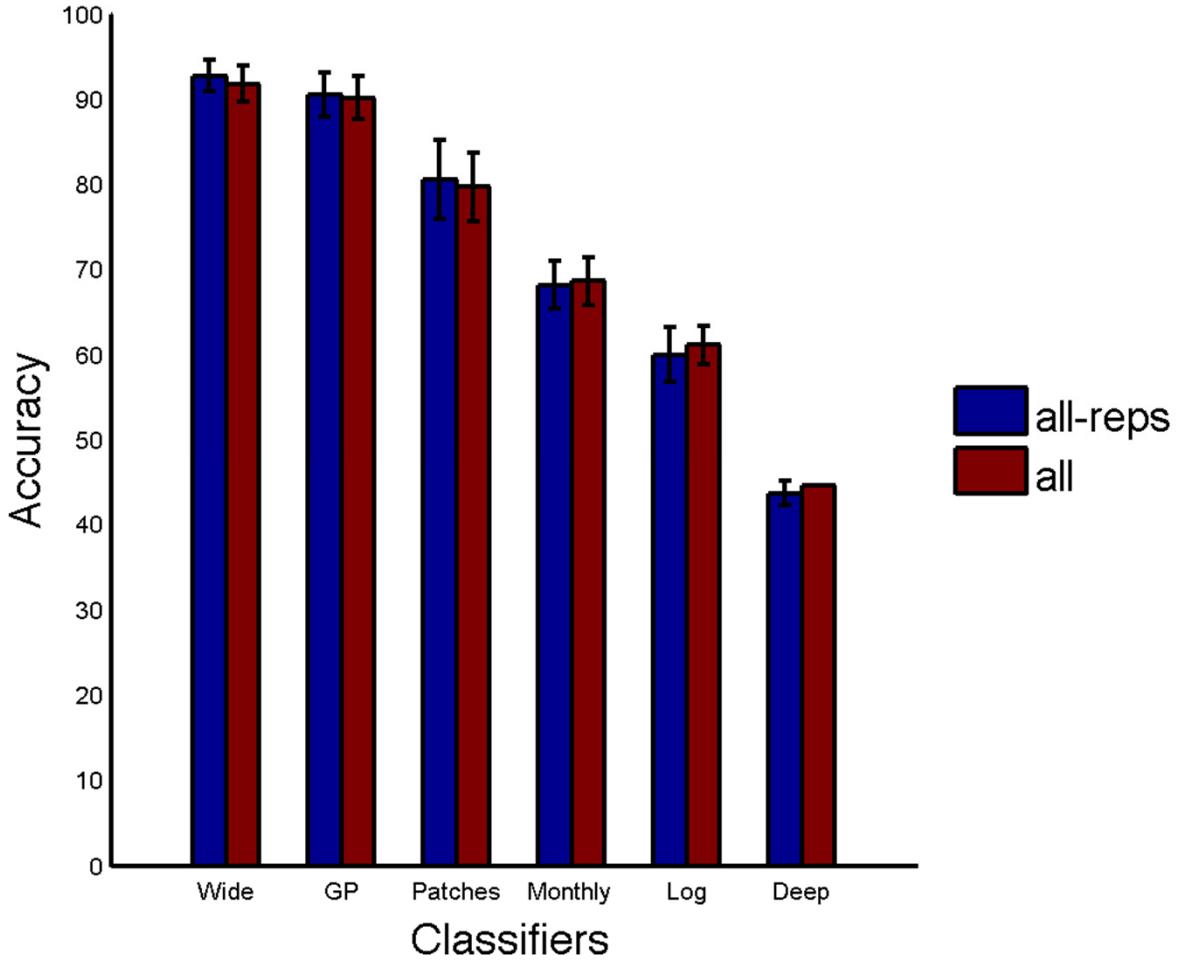


Accuracy by Classifier

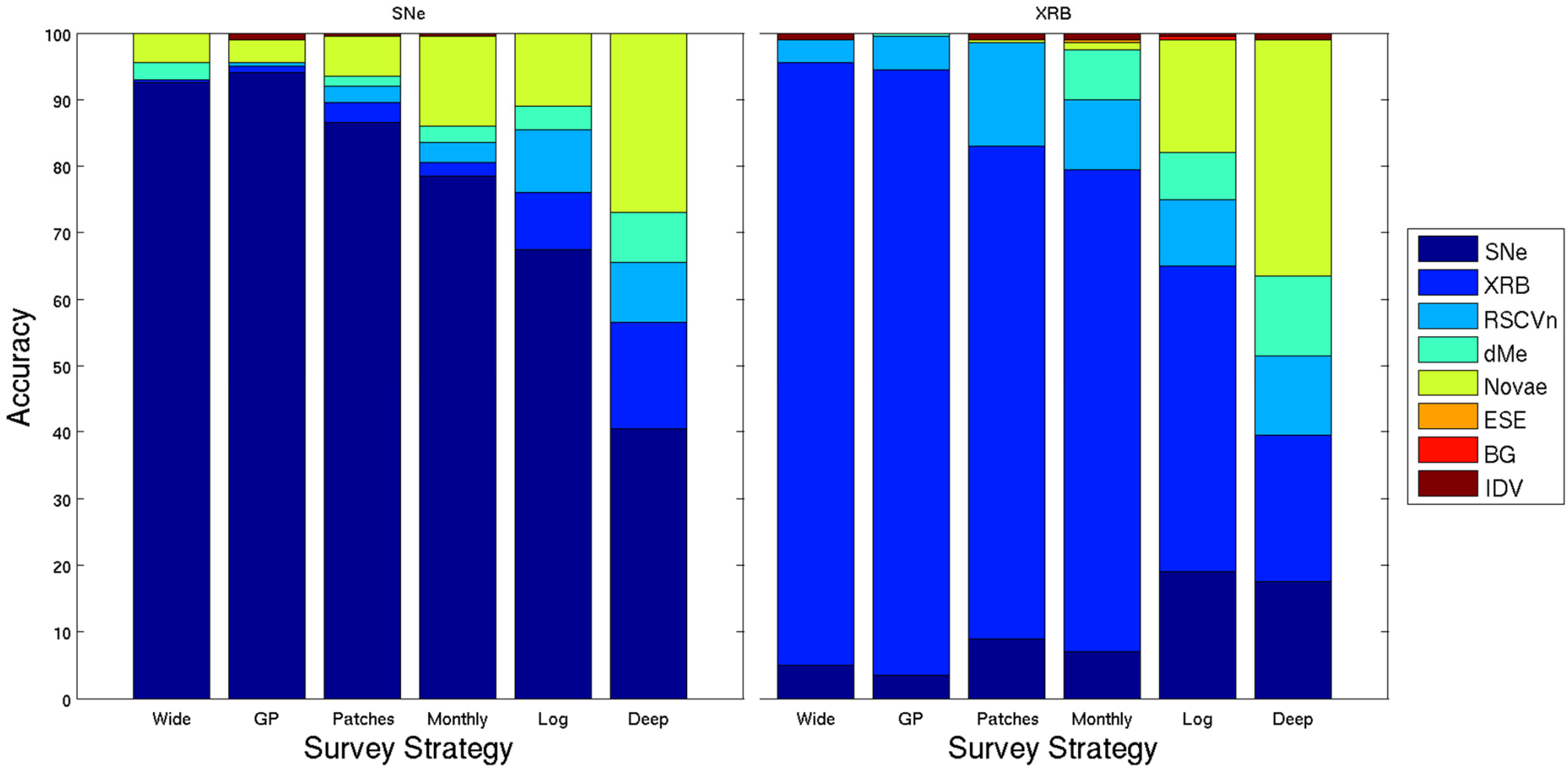


Accuracy by Survey Strategy

- VAST GP has similar performance to Wide with ¼ observations
- Early observations matter

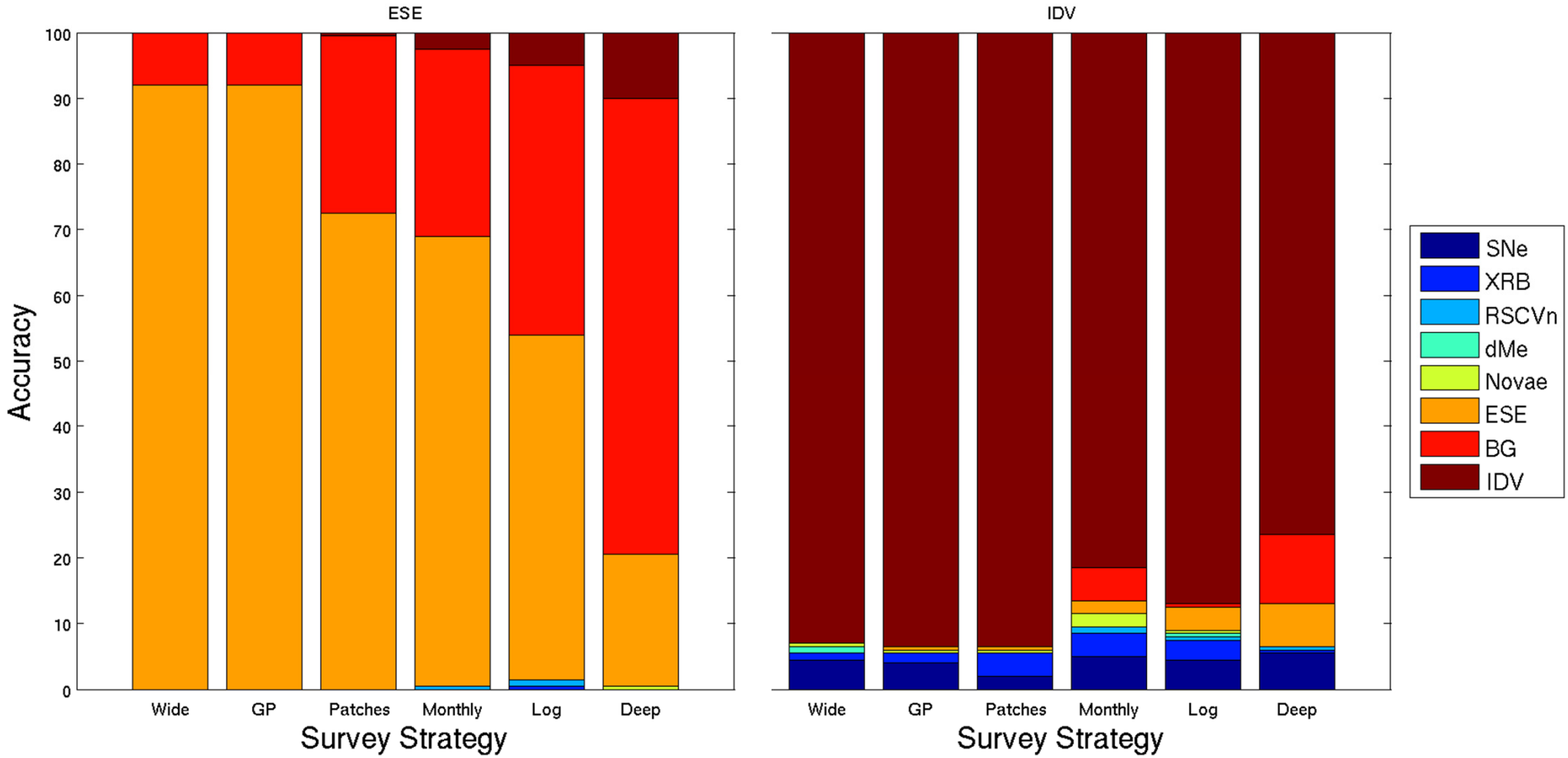


Class Confusions



SNe **XRB**

Class Confusions



ESE **IDV**

Archival: Conclusions

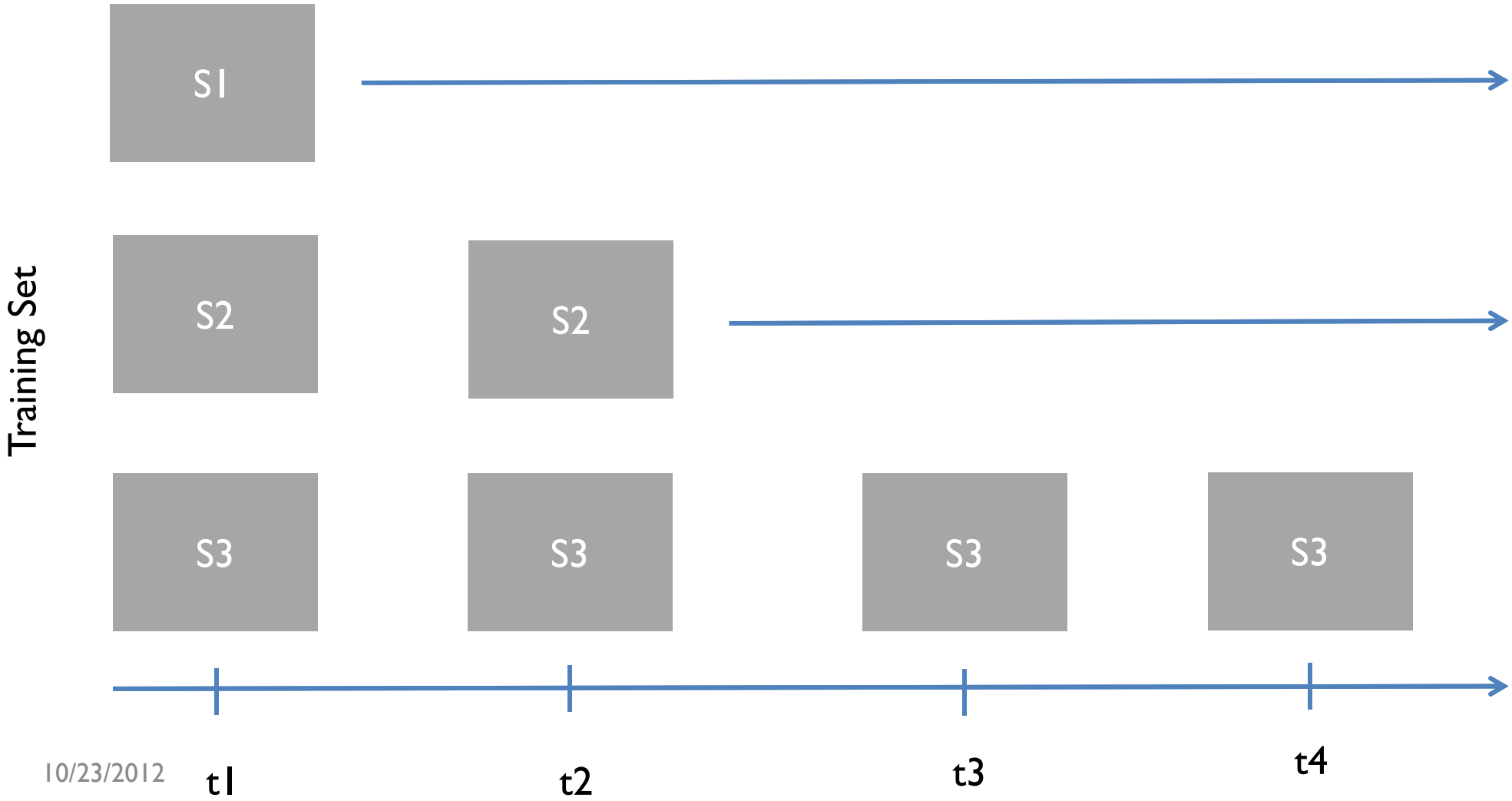
- Winner: SVM with cumulative statistics and concatenated features

- Two major confusion groups:
    - BG, IDVs, ESE
    - Sne, Novae, Flare Stars (RSCVn and dME), XRBs

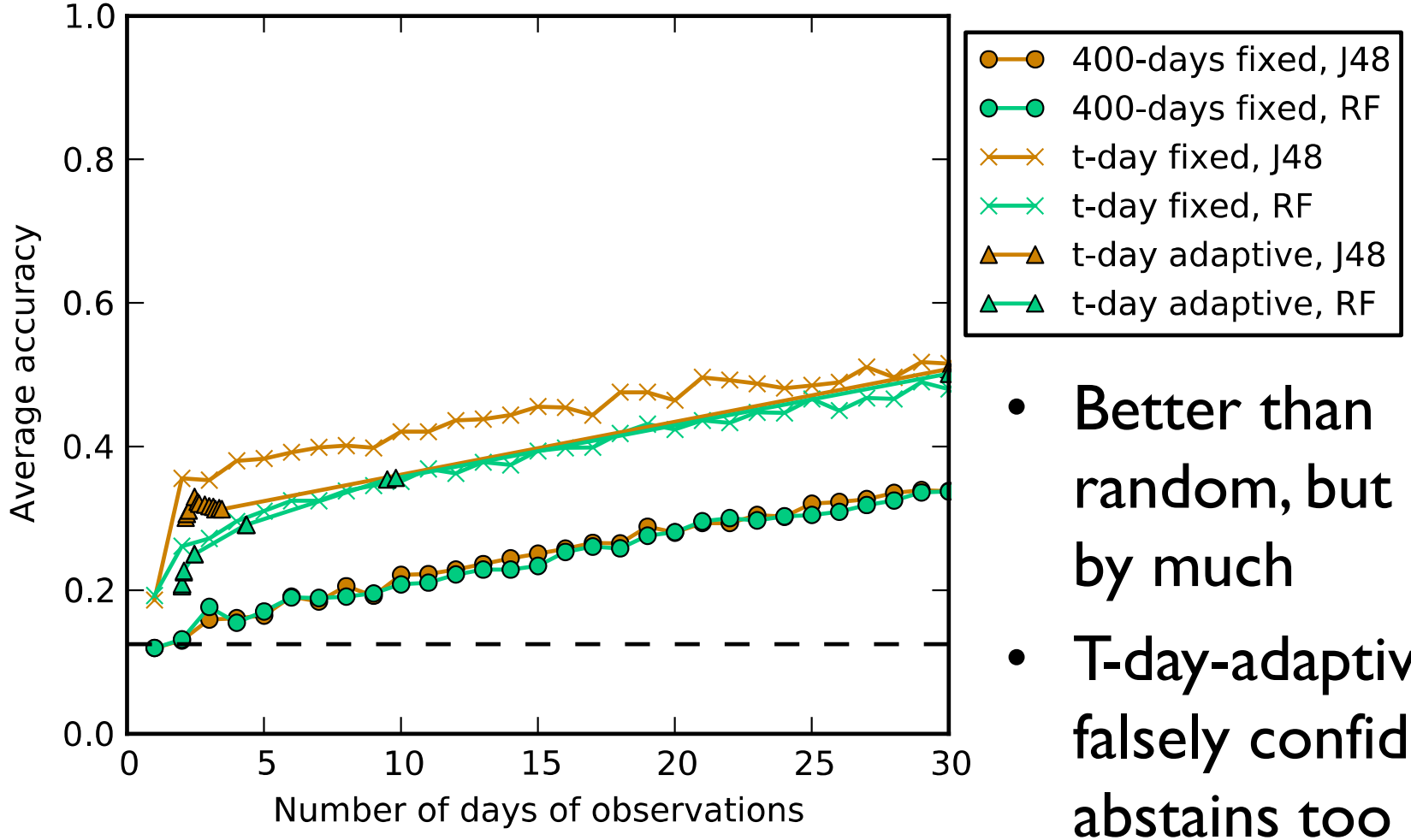
Online Classification

- 400-day-fixed
  - train on full archival knowledge, test on partially-observed light curve
  - No attempt to map train to test distribution
- t-day-fixed
  - Build ensemble of classifiers, each built only with t observations
  - Match test light curve with classifier built with same number of observations
- t-day-adaptive
  - Builds same ensemble
  - Outputs classification decision when confident of prediction, otherwise abstains and waits for more observations
  - Enables faster decision

t-day-adaptive

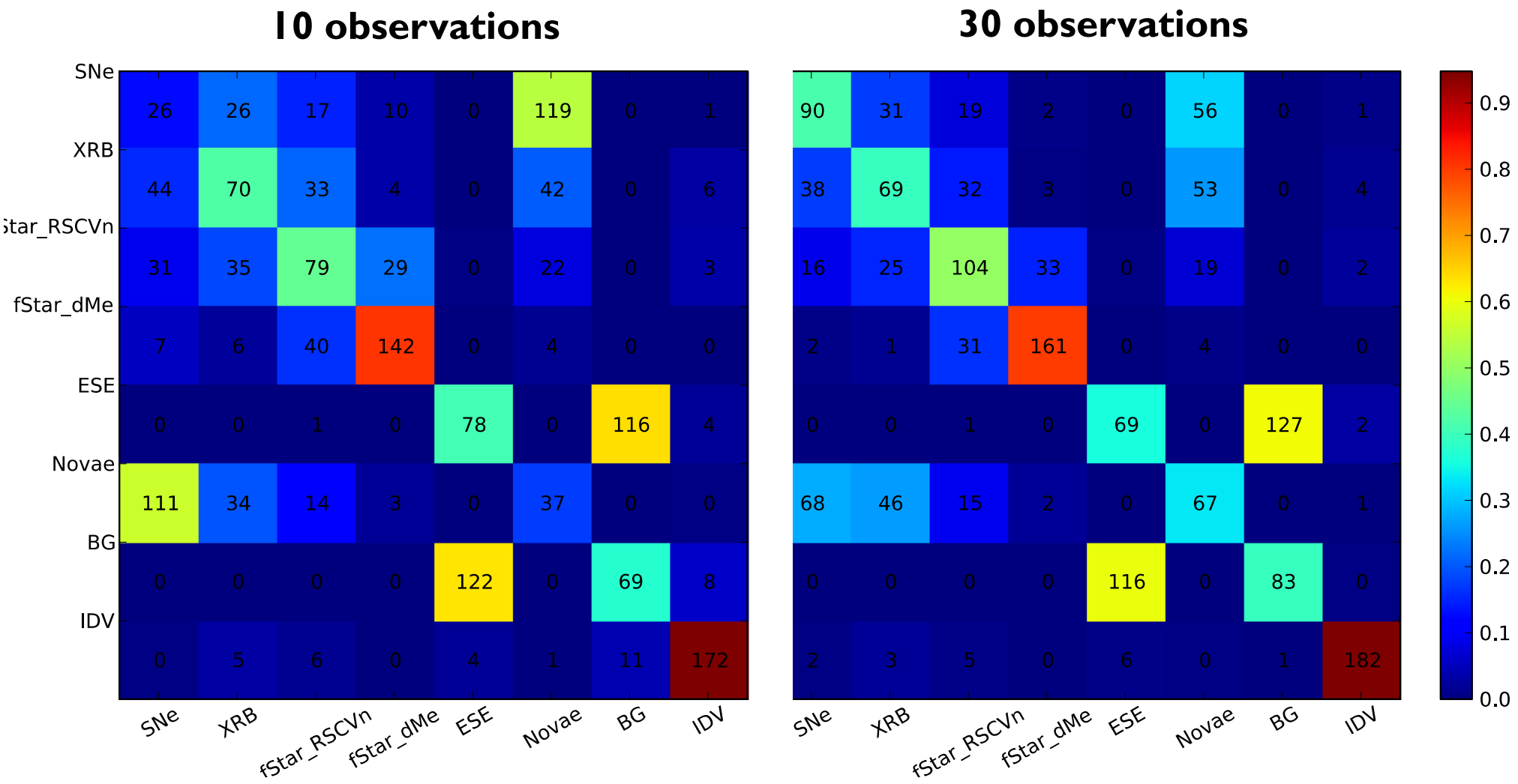


Online: Results

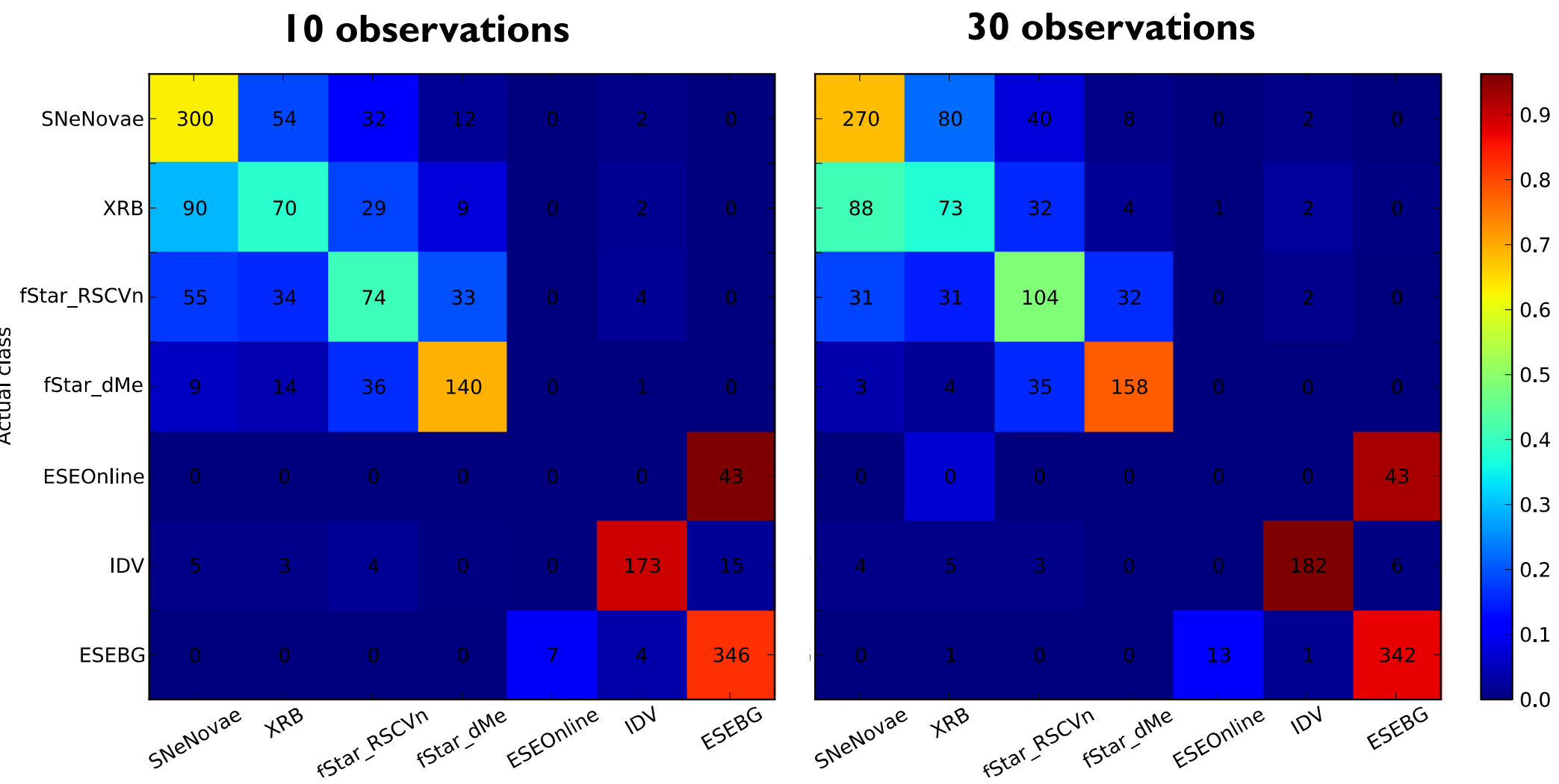


- Better than random, but not by much
- T-day-adaptive falsely confident, abstains too early

Online: Class Confusions



Online: Subgroup Ambiguous Classes



- Sne and Novae grouped together

Online: Subgroup Ambiguous Classes

- Regroup
  - Transients: SNe, XRB, fStar_RSCVn, fStar_dMe, Novae
  - Variables: IDV, ESEs, BG
- With 2 observations, accuracy = ~ 99%

	Pred. Transients	Pred. Variables
True Transients	990	5
True Variables	8	589

Future Work

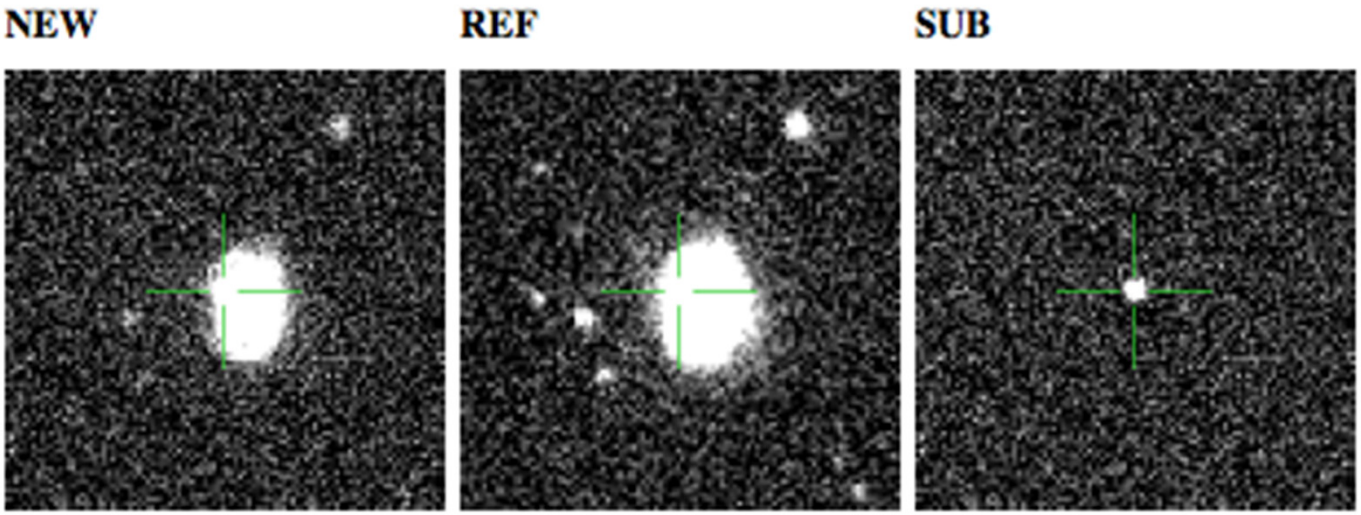
- Hierarchical classification approach
- Benchmark methods on optical data sets (MACHO, CRTS)
- Integration into VAST data processing pipeline

Agenda

- Overview of Research Interests
- ASKAP / VAST
  - Source type classification
  - Results of study on simulated data
  - Archival, online classification
- **Palomar Transient Factory**
  - Binary real time classification
  - Preliminary results

Palomar Transient Factory

- Candidates in subtracted images
- Classify in real time as "real" or "bogus"

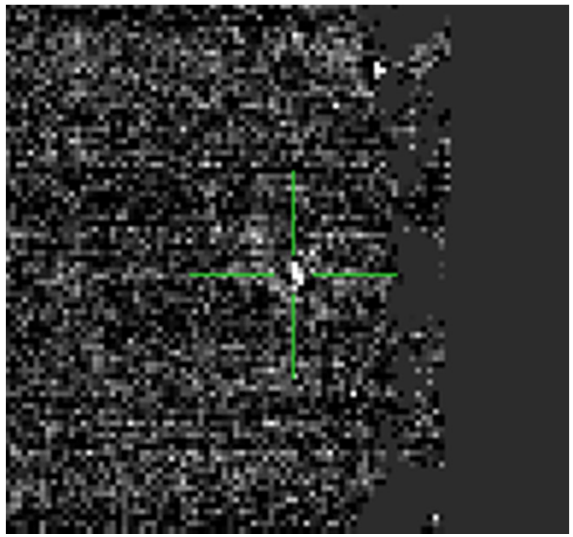


Sources of Error

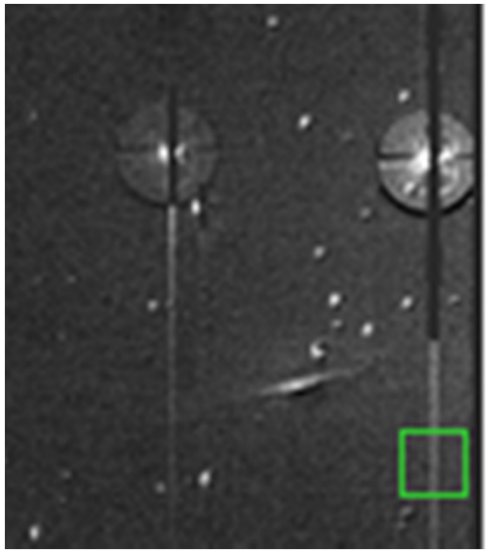
- Image subtraction process:
  - Bad PSF convolution
  - Pixel saturation
  - Diffraction spikes
  - Source close to edge
  - Bad alignment
- False positives
  - Cosmic rays
- In most cases, a human can easily identify a bogus candidate

Examples

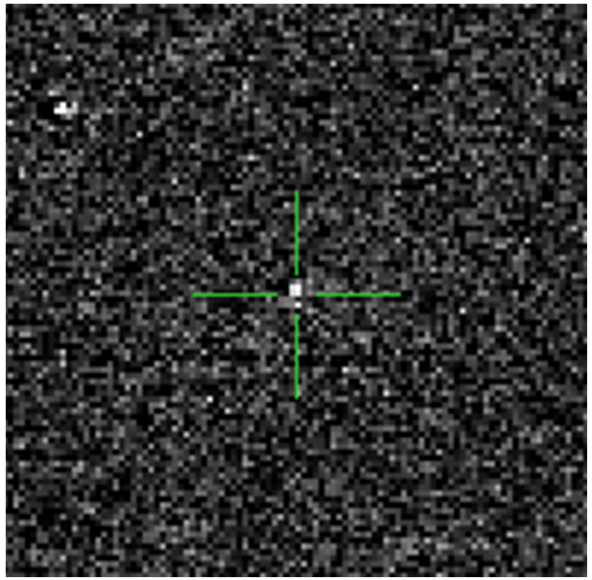
- In most cases, a human can identify the problem.



Junky PSF



Diffraction Spike



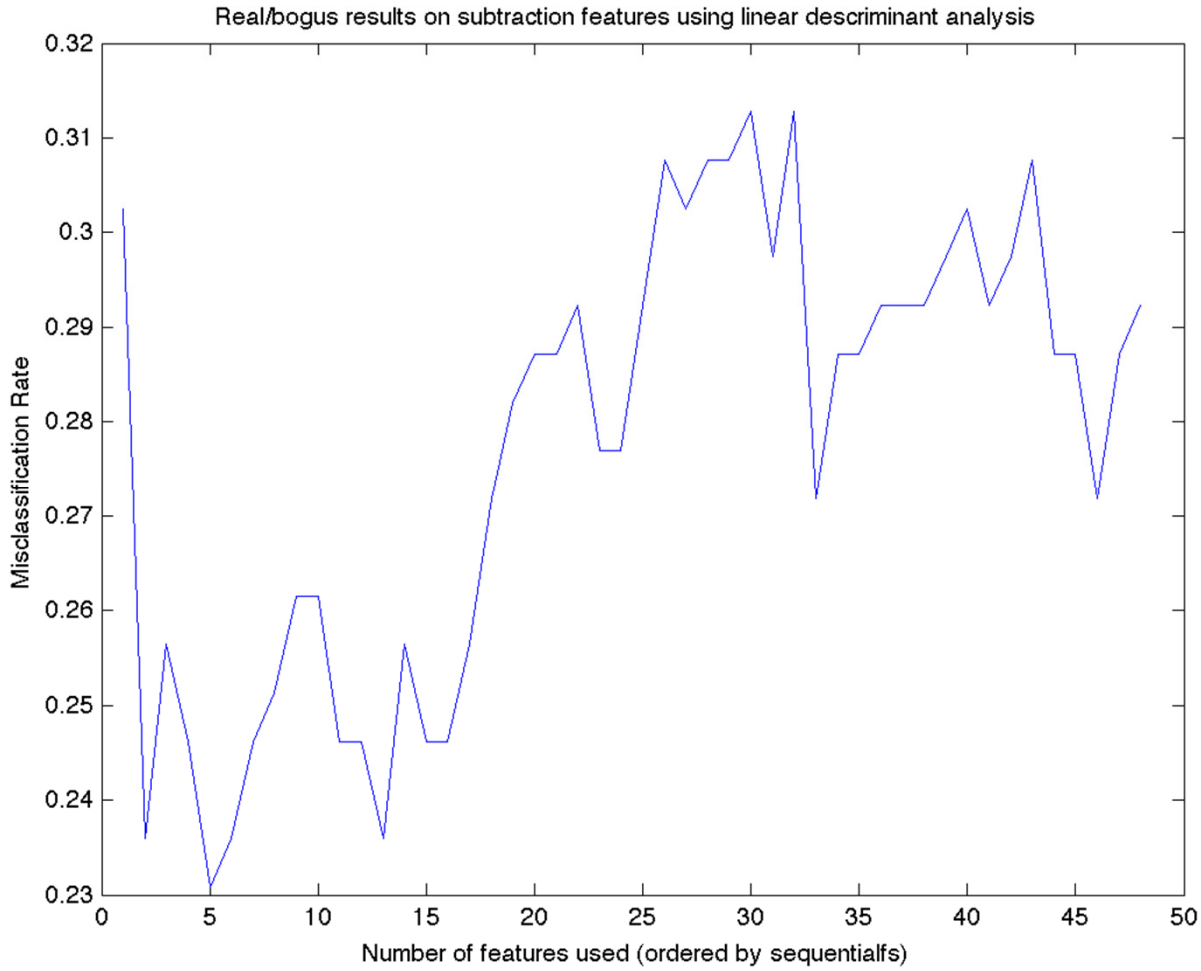
Bad Subtraction (neighbor in top left didn't subtract out)

Machine Learning Challenge

- How to encapsulate the human expert knowledge?
  - Feature that encapsulate that information
    - eccentricity of source
    - eccentricity nearest neighbors
  - Features from:
    - Neighboring sources
    - Time domain
- Adequate training data
  - reflects true distribution of real/bogus sources
  - Active learning?

Preliminary Results

- 195 Examples
  - 132 Bogus, 63 Real from local universe
  - Raw database features
  - Majority class classification: 32.3% error
  - Linear discriminant with sequential forward selection
  - 10-fold CV



Preliminary Results

- Random Forest

	Pred. Bogus	Pred. Real
True Bogus	116	16
True Real	28	35

- False positive rate: ~ 12%
- False negative rate: ~ 44%

Future Work

- Extract features from neighbors and time domain information
- Incorporate domain knowledge algorithmically
- Improving training set data